The Journal of Open Source Software

fairmetrics: An R package for group fairness evaluation

Benjamin Smith[1], **Jianhui Gao**[1], **Benson Chou**[1], and **Jessica Gronsbell**[1]

1 Department of Statistical Sciences, University of Toronto

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Fairness is a growing area of machine learning (ML) that focuses on ensuring that models do not produce systematically biased outcomes across groups defined by protected attributes, such as race, gender, or age. The `fairmetrics` R package provides a user-friendly framework for rigorously evaluating group-based fairness criteria, including independence (e.g., statistical parity), separation (e.g., equalized odds), and sufficiency (e.g., predictive parity) for binary protected attributes. The package provides both point and interval estimates for a variety of commonly used criteria. `fairmetrics` also includes an example dataset derived from the Medical Information Mart for Intensive Care, version II (MIMIC-II) database (Goldberger et al. 2000; J. Raffa 2016; J. D. Raffa et al. 2016) to demonstrate its use.

Statement of Need

ML models are increasingly used in high-stakes domains such as criminal justice, healthcare, finance, employment, and education (Mehrabi et al. 2021; Mattu 2016; Gao et al. 2024). Existing fairness evaluation software report point estimates and/or visualizations, without any measures of uncertainty. This limits users' ability to determine whether observed disparities are statistically significant. `fairmetrics` addresses this limitation by including confidence intervals (CIs) for both difference and ratio based fairness metrics to enable more robust and statistically grounded fairness assessments.

Fairness Criteria

`fairmetrics` is designed to evaluate fairness of binary classification models across binary protected attributes. The package supports the evaluation of metrics belonging to three major group fairness criteria:

- **Independence:** Statistical Parity (compares the overall rate of positive predictions between groups).
- **Separation:** Equal Opportunity (compares false negative rates between groups), Predictive Equality (compares false positive rates between groups), Balance for Positive Class (compares the average predicted probabilities among individuals whose true outcome is positive across groups), and Balance for Negative Class (compares the average predicted probabilities among individuals whose true outcome is negative across groups).

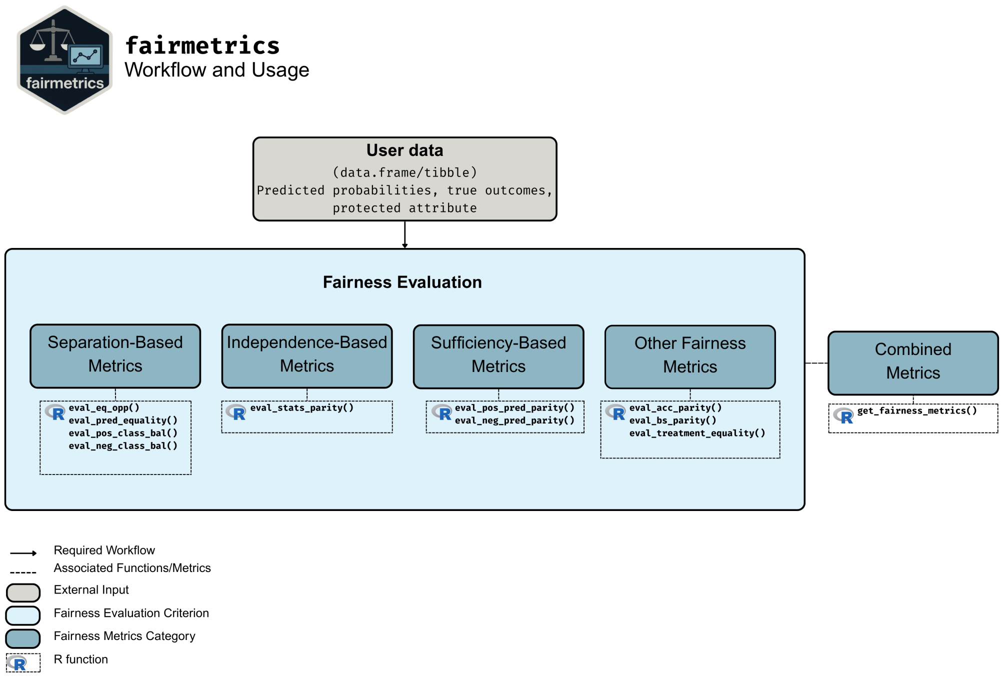

Figure 1: Workflow for using `fairmetrics` to evaluate model fairness across multiple criteria.

- **Sufficiency:** Positive Predictive Parity (compares the positive predictive values across groups), Negative Predictive Parity (compares the negative predictive values across groups).

The package also includes additional metrics, such as the Brier Score Parity (compares the Brier score across groups), Accuracy Parity (compares the overall accuracy across groups), and Treatment Equality (compares the ratio of false negatives to false positives across groups).

Evaluating Fairness Criteria

The input required to evaluate model fairness with the `fairmetrics` package is a `data.frame` or `tibble` containing the model's predicted probabilities, the true outcomes, and the protected attribute. Figure 1 shows the workflow for using `fairmetrics`.

A simple example of how to use the `fairmetrics` package is illustrated below. The example makes use of the `mimic_preprocessed` dataset, a pre-processed version of the Indwelling Arterial Catheter (IAC) Clinical dataset, from the MIMIC-II clinical database (Goldberger et al. 2000; J. Raffa 2016; J. D. Raffa et al. 2016).

While the choice of fairness metric used is context dependent, we show all criteria available with the `get_fairness_metrics()` function for illustrative purposes. In this example, we evaluate the model's fairness with respect to the binary protected attribute `gender`. The model is trained on a subset of the data and the predictions are made and evaluated on a test set. A statistically significant difference across groups at a given level of significance is indicated when the CI for a difference-based metric does not include zero or when the interval for a ratio-based metric does not include one.

```r
library(fairmetrics)
# Setting alpha=0.05 for 95% CIs
get_fairness_metrics(
 data = test_data,
 outcome = "day_28_flg",
 group = "gender",
 probs = "pred",
 cutoff = 0.41,
 alpha = 0.05
)

           Fairness Assessment                             Metric
1          Statistical Parity           Positive Prediction Rate
2           Equal Opportunity                False Negative Rate
3         Predictive Equality                False Positive Rate
4  Balance for Positive Class      Avg. Predicted Positive Prob.
5  Balance for Negative Class      Avg. Predicted Negative Prob.
6  Positive Predictive Parity          Positive Predictive Value
7  Negative Predictive Parity          Negative Predictive Value
8          Brier Score Parity                        Brier Score
9     Overall Accuracy Parity                           Accuracy
10         Treatment Equality (False Negative)/(False Positive) Ratio

   GroupFemale GroupMale Difference    95% Diff CI Ratio 95% Ratio CI
1         0.17      0.08       0.09   [0.05, 0.13]  2.12 [1.49, 3.04]
2         0.38      0.62      -0.24 [-0.39, -0.09]  0.61 [0.44, 0.86]
3         0.08      0.03       0.05   [0.02, 0.08]  2.67  [1.4, 5.08]
4         0.46      0.37       0.09   [0.04, 0.14]  1.24 [1.09, 1.42]
5         0.15      0.10       0.05   [0.03, 0.07]  1.50 [1.29, 1.74]
6         0.62      0.66      -0.04  [-0.21, 0.13]  0.94 [0.72, 1.22]
7         0.92      0.90       0.02  [-0.02, 0.06]  1.02 [0.98, 1.07]
8         0.09      0.08       0.01  [-0.01, 0.03]  1.12 [0.89, 1.43]
9         0.87      0.88      -0.01  [-0.05, 0.03]  0.99 [0.94, 1.04]
10        1.03      3.24      -2.21 [-4.38, -0.04]  0.32 [0.15, 0.68]
```

Users can also compute individual metrics using functions like `eval_eq_opp()` to test specific fairness conditions. Full usage examples are provided in the package documentation.

Related Work

Other R packages similar to `fairmetrics` include `fairness` (Kozodoi and V. Varga 2021), `fairmodels` (Wiśniewski and Biecek 2022) and `mlr3fairness` (Pfisterer, Siyi, and Lang 2024). `fairmetrics` differs from these packages in two ways. The first difference is that `fairmetrics` calculates ratio and difference-based group fairness metrics and their corresponding CIs, allowing for more meaningful inferences about the fairness criteria. The second difference is that `fairmetrics` does not possess any external dependencies and has a lower memory footprint. Table 1 shows the comparison of memory used and dependencies required when loading each library.

For Python users, the `fairlearn` library (Weerts et al. 2023) provides additional fairness metrics and algorithms. The `fairmetrics` package is designed for seamless integration with R workflows, making it a more convenient choice for R users.

Package	Memory (MB)	Dependencies
fairmodels	17.02	29
fairness	117.61	141
mlr3fairness	58.11	45
fairmetrics	0.05	0

Table 1: Memory usage (in MB) and dependencies of 'fairmetrics' vs similar packages.

Licensing and Availability

The `fairmetrics` package is under the MIT license. It is available on CRAN and can be installed by using `install.packages("fairmetrics")`. Full documentation and its examples are available at: https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html. Source code and issue tracking are hosted on GitHub: https://github.com/jianhuig/fairmetrics/.

References

Gao, Jianhui, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. 2024. "What Is Fair? Defining Fairness in Machine Learning for Health." *arXiv.org*. https://doi.org/10.48550/arXiv.2406.09307.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation [Online]* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *fairness: Algorithmic Fairness Metrics*. https://doi.org/10.32614/cran.package.fairness.

Mattu, Lauren Kirchner, Jeff Larson. 2016. "Machine Bias." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Comput. Surv.* 54 (6). https://doi.org/10.1145/3457607.

Pfisterer, Florian, Wei Siyi, and Michel Lang. 2024. *mlr3fairness: Fairness Auditing and Debiasing for 'Mlr3'*. https://doi.org/10.32614/cran.package.mlr3fairness.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F.

Raffa, Jesse D., Mohammad Ghassemi, Tristan Naumann, Mengling Feng, and Daniel J. Hsu. 2016. "Data Analysis." In *Secondary Analysis of Electronic Health Records*, 109–22. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_9.

Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. "FairLearn: Assessing and Improving Fairness of AI Systems." *arXiv.org*. https://doi.org/10.48550/arXiv.2303.16626.

Wiśniewski, Jakub, and Przemysław Biecek. 2022. "fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models." *The R Journal* 14 (1): 227–43. https://doi.org/10.32614/RJ-2022-019.